Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 1088 – 1095

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Training for the unknown: The role of feedback and similarity in detecting zero-day attacks

Noam Ben-Asher[a,b,*], Cleotilde Gonzalez[c]

[a]*Army Research Laboratory, Adelphi, MD, USA*
[b]*IBM T.J.Watson Research Center, Yorktown Heights, NY, USA*
[c]*Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University, PA, USA*

Abstract

Human cognitive and analytical capabilities are needed and are indispensable to success in cyber defense. However, the high volume of network data challenges the process of detecting cyber-attacks, especially zero-day attacks. Training along with detailed and timely outcome feedback is a major factor in improving performance. It supports attributes identification and rule formation, which are crucial to the detection of attacks. To understand the role of feedback during training and how it influences the detection of novel attacks, we developed a simplified Intrusion Detection System and trained 160 participants to perform as analysts. During training, participants classified network events representing a specific cyber-attack, and received feedback at the end of each trial. Detailed feedback used color schemes informing of hits, misses, false-alarms, and correct-rejections. Aggregated feedback provided numerical summaries regarding performance. After training, participants classified events that were similar or part of a novel attack. Results show that detailed feedback accelerated learning and improved detection accuracy compared to aggregated feedback. Participants who received aggregated feedback failed to learn the role of certain network attributes and how to integrate them into detection rules. Surprisingly, aggregated feedback improved detection in the novel attack. The novelty of a situation caused an increase in decision scrutiny, while familiar decision situations limited the depth of information search and evaluation. Analyst should learn to abstract information and look broadly at outcome feedback to improve their ability to detect novel attacks. We discuss the implications of these findings for improving cyber security.

Peer-review under responsibility of AHFE Conference

Keywords: Learning; Feedback; Similarity; Cybersecurity; Cyberattack

* Corresponding author.
E-mail address: nbenash@us.ibm.com

Peer-review under responsibility of AHFE Conference
doi:10.1016/j.promfg.2015.07.180

1. Introduction

Disrupting computers and the loss of sensitive information through cyber-attacks are becoming a widespread threat and a critical concern. Guarding against such attacks is a significant part of network governance done by cyber security analysts [1]. Even with recent advances in information and network security and the development of new monitoring and threat detection tools, the tasks performed by the analyst cannot be completely automated. The analytical capabilities of the human decision maker are needed and indispensable [2]. In their work, security analysts heavily rely on Intrusion Detection System (IDS), technologies that can detect network intrusions and network misuse by matching patterns of known attacks against ongoing network activity. Once the IDS finds a match to a known type of attack or detects abnormal network activity, it produces alerts detailing the suspicious events [3]. Considering the amount of traffic in a mid-size corporate network and the ever-growing number and complexity of cyber-attacks, the number of alerts generated can be overwhelming. Such systems can trigger thousands of alerts per day, up to 99% of which are false alerts [3]. Eventually, the high volume of intrusion alerts that needs to be processed and the high probability of false alerts make the process of accurately detecting a cyber-attack challenging for human cognitive capabilities.

There is a growing body of work within the cyber security field that focuses on understanding the work processes of analysts [3, 4, 5]. Monitoring and detection belong to a general process called triage analysis, which fundamentally depend on the analyst's knowledge [6]. Usually, analysts undergo extensive training and certification programs to attain such knowledge and expertise. However, despite the cardinal role of training, little is known about how to train analysts to detect threats in a highly dynamic cyber space. Even less is known about how to support the detection of zero-day attacks, which are novel attacks that take advantage of unknown vulnerabilities, or how analysts generalize existing knowledge to judge and make a decision regarding a novel network situation.

Understanding how the analyst detects relevant attributes and constructs the appropriate decision rule to determine the network's status is extremely important. If critical attributes are overlooked, an attack can go unnoticed and cause severe damage. Similarly, agile rule formation supports timely detection and can assist in the detection of novel threats. Proper training, along with detailed and timely outcome feedback, is a major factor in improving performance [7]. Such feedback can support both attributes identification and rule formation. This study examines how training with feedback, which does not explicitly point to relevant attributes or to actual rules, influences performance when confronting familiar or novel cyber-attacks. This implies that it is possible to increase the consideration of an attribute as relevant and its inclusion in a classification rule.

2. Attribute identification and rule learning in cyber security

Network traffic data is big and complex because many records are generated per second and each have many attributes. For example, the description of network traffic can include timestamp, protocol type, source IP, destination IP, port activity, and duration among others. Considering the quantity and complexity of network data, it is important to provide the analyst with the right information that allows fast and accurate detection. Providing sparse information or overwhelming amounts of information can undermine the analyst's judgments and decrease detection accuracy. Furthermore, a single attribute cannot usually provide sufficient indication of a network's status. Thus, the analyst has to consider relatively complex rules that combine multiple attributes. Such rules accommodate the existence of internal relationships between the attributes and the exact internal relationship. For example, high network activity between two internal IP addresses is treated differently from high network activity from an internal IP address to an external one. In this example, three attributes (i.e., source IP, destination IP, and the amount of traffic between them) combine to form a rule that determines the type of a network activity.

Such decision making involves an iterative process of identifying relevant attributes and discovering a conceptual rule. Attributes identification is a preliminary requirement for rule formation, where a subset of relevant attributes is attended to from a larger set of attributes. The number of relevant and irrelevant attributes, the number of values each attribute can take, the variability within these values, and redundancy or overlap between attributes can impact identification [8, 9]. According to a relevant set of attributes, a rule can be formed. Rules are considered the atoms of thought, and as such, they are of great interest to cognitive science [10]. Some rules are easy to discover and learn, while others are more difficult and less intuitive [11]. Recent approach to rule complexity uses Boolean

algebra and logical operations that combine attributes as a measure of complexity. Hence, the subjective difficulty of a rule is proportional to the minimal length of the Boolean formula that represents the rule [11].

Through repeated interaction, it is possible to learn a rule from experience and successfully apply it in a specific environment—this increases performance in terms of accuracy and speed. Furthermore, rules also support abstract reasoning and are crucial for reasoning across contexts. They allow for the transfer of abstract knowledge learned in one environment (i.e., training environment) to other environments (i.e., transfer environment). Rules devised in one environment can apply to and prove to be useful in other environments, and therefore improve performance in novel situations. The extent to which it is possible to transfer a rule between environments depends on the similarity between the environments [10, 12].

Similarity takes many forms and definitions. Deep similarity is described as structural consistency with one-to-one mapping between low level elements and as consistency in the relationships between the elements. In contrast, surface similarity usually refers to perceptual similarity, which is a result of a high level comparison of aggregated attributes [12]. For example, novices rely more on surface similarity while experts tend to rely on structural similarity [13]. In many real world situations, however, generalization and learning can benefit from both deep and surface similarity as long as there is no conflict between them.

Similarity is used in cyber security as part of alert correlation algorithms. It helps identify common attack patterns, which occur at different points in the network or were previously observed on different networks [14]. The Diamond Model of intrusion analysis [15] illustrates the importance of similarity in identifying malicious events and common network behaviors. The model describes how the use of similarities in infrastructure and cyber capabilities support attribution of attacks and how similarity can be utilized to predict future targets of an attacker. Similarity is also important to the detection of zero-day attacks when operating on the assumption that there is some similarity between the attack vector in use and previously detected attack vectors in zero-day exploits [16]. Despite the use of similarity in cyber security, there is limited understanding of how analysts detect similarities between previously observed attacks and novel attacks. Moreover, little is known about how analyst can be trained to utilize similarity when attending to the relevant attributes and forming an appropriate classification rule. To better understand the interplay between the type of feedback during training, detection of threats, and similarity, we developed a simplified IDS. Our hypothesis is that detailed feedback during training will have a positive effect on threat detection. However, training with aggregated feedback, which hints at the correct classification without surrendering all the details, might has positive effect on the detection of novel attacks.

3. Simplified intrusion detection task

In this study, participants served as security analysts of a fictitious online retail company. Their duty was to protect the company's network from a malicious attacker. Based on the network described by Lye and Wing [17], illustrated in Figure 1, we used a simple stereotypical computer network. This network setting is commonly used in

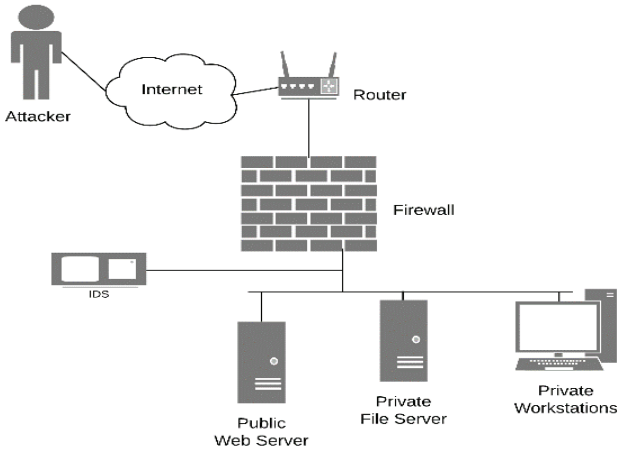

Fig. 1. The local computer network.

cyber security research and training, as well as in the operative networks of real-world mid-size corporations [6, 17, 18]. Such networks typically consist of a web server, a file server, and a cluster of workstations.

Detailed instructions described the network nodes and the dependencies between them (for a detailed description, see [6]). The security analyst (i.e., the participant) monitors this network by observing sequences of network events through the interface seen in Figure 2. Each sequence includes network events that represents a single independent network scenario. The order of the events within a scenario corresponds to the order in which they occurred over the network, and the participant's goal is to decide whether or not a network event is malicious. Each network event is presented one by one and has an identification number corresponding to its presentation. The event can also include an Alert, which provides information regarding a suspicious network activity; for example, indicating that a service has stopped or started. The IDS can generate false alerts and can also miss malicious events. The last part of the display presents the event's description. The structure of the description is consistent, and it includes the type of network node (e.g., web server) and the processes which are currently running on it (e.g., httpd), the traffic load going between that node and another node (e.g., between the web server and the fileserver). Traffic load had one of the following values: 0Mbps, 3.3Mbps, 6.7Mbps, and 10Mbps; where 3.3Mbps is the normal traffic condition, 0Mbps indicates that there is no traffic between two nodes, and 10Mbps indicates maximal capacity. Additional information in the description indicates whether or not an operation was executed on this network node. The simplified information for each network event resembled the presentation of events' signature in IDS, with simplifications allowing participants without experience with IDS to comprehend it. An analyst can classify each event as malicious or not by checking or un-checking the corresponding "Is threat" box for each event. Upon completing a network scenario, participants received feedback and a message informing them that a new network scenario is about to appear.

4. Method

4.1. Design and procedure

After classifying all the events in a scenario, the experimental system provided feedback regarding the classification. We evaluated two types of feedback and their influence on learning and the ability to generalize the learning process to novel scenarios. A detailed feedback, illustrated Figure 2, used a color scheme to inform the participants of the hits (bright green), misses (bright red), false-alarms (orange), and correct-rejections (light green) they made in a trial. This information was accompanied with numeric summaries of the current and total earnings based upon performance. In the aggregated feedback conditions, participants received summarized numerical information regarding their performance in each trial. The feedback included the number of correctly classified

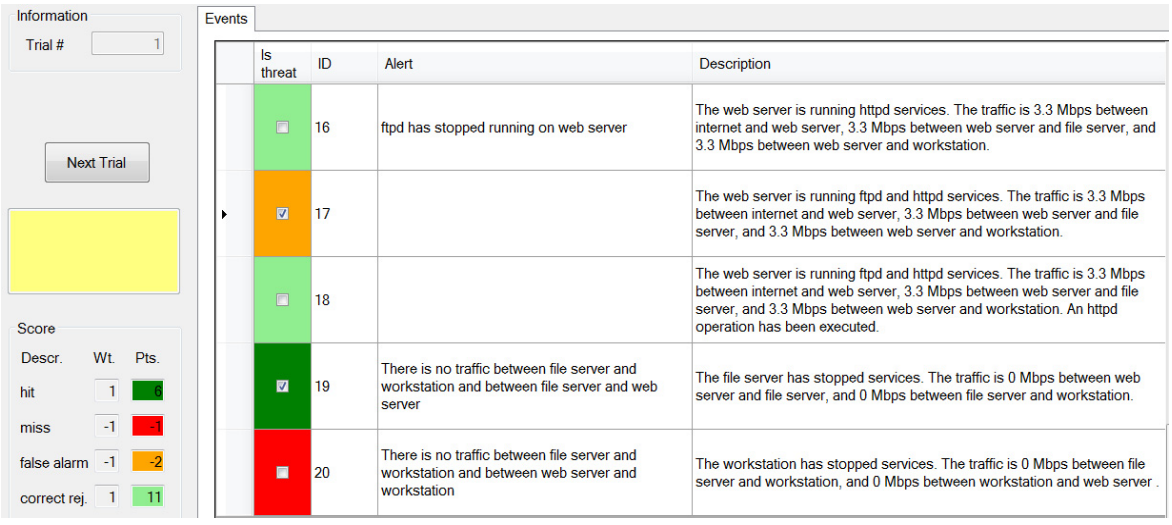

Fig. 2. The IDS displaying detailed feedback after classification of network events.

events in the trial without separately categorizing hits, misses, false-alarms, and correct-rejections. The feedback also provided the current and total earnings based upon performance. Unlike the detailed condition, however, participants did not see which events were classified correctly.

The experiment involved a 2×2 factorial design, with feedback type (Aggregated or Detailed) and type of transfer environment (Similar or Novel) as the independent variables. Hence, there were four experimental conditions, each with 40 participants. The experiment included two phases: training and transfer. There is a training phase that consisted of eight consecutive trials, during which participants classified network events in eight versions of the same network scenario and received detailed or aggregated feedback. When the training phase ended, participants were informed they would now do the transfer phase, where participants classified network events in two network scenarios without receiving any feedback. In the Similar conditions, participants classified events in two network scenarios that represented the same cyber-attack as in the training phase. While in the Novel conditions, participants classified events in two scenarios that represented a novel attack type not previously observed during training.

4.2. Network scenarios

An attacker outside the corporation may try to gain access to the corporate network to obtain confidential information or to compromise an essential service. For this, the attacker follows an "island-hopping" attack [2], where the web server is compromised first and is then used to originate attacks on the file server or on the company workstations. Based on the network structure illustrated in Figure 1, we defined two network scenarios; each represents a different type of an attack. We used the same ground truth rule for both network scenarios. An event is *malicious* if there is an alert, and if the description of the event indicates irregular network traffic or that an operation was executed on one of the network node (or both). Thus, a network event is malicious if it follows this rule: Alert $\cap$ (Operation $\cup$ Network Load). Each scenario was composed of 20 network events, with a base-rate of .35 malicious events. As the IDS can generate false alerts, it generated three false alerts in each of the scenarios.

For training, we used a network scenario representing intrusion and data exfiltration. This scenario starts with normal network operation and ends after the attacker attained confidential information and disrupted the network's regular operation. The attacker gains access to the web server by compromising one of the services it runs. Once the attacker has hacked into the web server, he installs a backdoor program that provides access to the file server. Then the attacker steals confidential information stored on the file server and causing even more damage to the network, the attacker shuts the network down using the obtained privileges. We generated a pool of eight versions of this scenario by applying changes to the content and order of the network events that are not related to the cyber-attack.

The transfer scenario represented intrusion and the installation of a password sniffer. In this scenario, the attacker gains control over the web server and installs a password sniffer. The sniffer collects the passwords of legitimate users as they access the file server or web server. Later, the attacker comes back and collects the password list from the sniffer. If the attacker goes unnoticed, the consequences can be disastrous. The attacker may have access to valuable information and network resources. This sniffer is detected by the IDS before the attacker manages to sabotage the network services.

4.3. Participants

We recruited 160 participants (46% women, $mean_{age}$=23.04, SD_{age}=2.90) to a computer laboratory, each was randomly assigned to one of the four experimental conditions. Participants were compensated with $10 as base payment and could earn additional monetary incentive based on performance. During the training phase, participants could earn 1 cent for each threat and non-threat correctly classified and lose 1 cent for each threat and non-threat incorrectly classified. During transfer, participants could earn 10 cents for each threat and non-threat correctly classified and lose 10 cents for each threat and non-threat incorrectly classified.

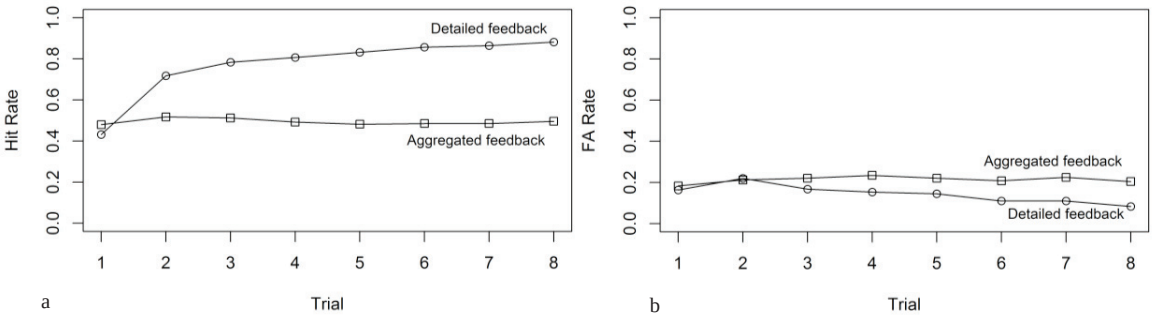

Fig. 3. Average (a) hit and (b) false alarm rate when detecting malicious network events in a trial for the Detailed and Aggregated feedback conditions.

5. Results

First, we analyze the performance in intrusion detection during the training and transfer periods. Then, we look closely at the network events themselves and evaluate how the attributes and the structure of the event influenced classification. As the participants' decisions in the first training trial were not influenced by feedback, the statistical analysis starts from the second training trial on; for completeness though, the plots show all data, including the first trial.

5.1. Training

5.1.1. Threat detection

Overall, we find that detailed feedback facilitated detection compared to aggregated feedback. Results shows higher hit rate in the Detailed condition (mean=.82, SD=.19) compared to the Aggregated condition (mean=.50, SD=.26), $F(1,158)=146.603$, $p<.001$. Similarly, the false-alarms rate in the Detailed condition (mean=.14, SD=.13) was lower than in the Aggregated condition (mean=.22, SD=.15), $F(1,158)=19.550$, $p<.001$. Combining these two measures of detection, using d-prime (d'), suggests that participants in the Detailed condition (d'=1.99) preformed the detection task better than participants in the Aggregated condition (d'=0.77).

As seen in Figure 3a, from the second trial onwards, there is an advantage to the Detailed feedback compared to the Aggregated feedback, as expressed by a higher hit rate. The significant two-way interaction between the feedback condition and trial indicates that in the Detailed condition performance in the Detailed condition improved with experience, while the hit rate remained relatively stable across trials in the Aggregated condition, $F(1,158)=24.049$, $p<.001$.

Analysis of the false-alarms rate also revealed a significant two-way interaction between feedback and trial, $F(1,158)=21.648$, $p<.001$. In the Detailed condition the false-alarms rate decreased with experience (see Figure 3b). In the Aggregated condition there was no decrease or increase in the false-alarms rates and thus, trial to trial learning did not improve participants' performance.

5.1.2. Attributes and rule structure

Participants in the Detailed condition detected more threats correctly compared to participants in the Aggregated for the threats in the form of *Alert+Load+Operation*, *Alert+Load*, and *Alert+Operation*, $t(158)=4.600$, $p<.001$; $t(158)=9.339$, $p<.001$ and $t(158)=11.478$, $p<.001$, respectively. This finding is consistent with the overall higher detection rate of participants in the Detailed condition. Analysing how each type of feedback influenced the detection of the different types of events, we find that in the Aggregated feedback condition, threats in the form of *Alert+Load* were correctly detected less often compared to threats in the form of *Alert+Load+Operation*, $t(158)=4.247$, $p<.001$. Furthermore, threats in the form of *Alert+Operation* were detected less successfully compared to threats in the form of *Alert+Load*, $t(158)=6.989$, $p<.001$. In contrast, in the Detailed condition, we find no significant differences in the detection of threats in the form of *Alert+Load+Operation* compared to *Alert+Load*, $t(158)=1.526$, p=ns. However, the detection of threats in the form of *Alert+Load* was higher compare to threats in the form of *Alert+Operation*, $t(158)=7.656$, $p<.001$.

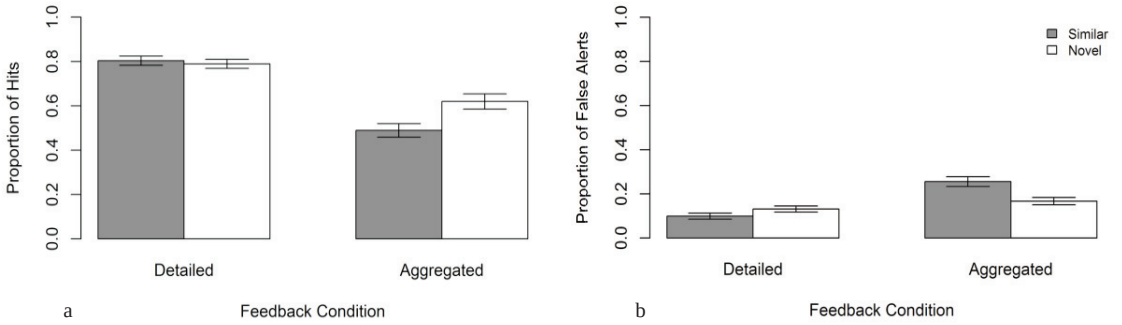

Fig. 4. Hits (a) and false alarm (b) rates during the transfer phase as a function of the event type and the feedback condition.

Analysis of the false alerts rates in each of the non-threat event types reveals that the differences in the false alert rate between participants in Detailed and Aggregated conditions also depended on the type of attributes that composed the event. Misclassification of events in the form of *Load Only* and *Load+Operation* were significantly lower in the Detailed condition compared with the Aggregated condition, $t(158)=4.899$, $p<.001$ and $t(158)=4.144$, $p<.001$, respectively. However, feedback condition had no effect on the false detection of events in the form of *Operation Only*, $t(158)=.862$, p=ns.

5.2. Transfer

5.2.1. Threat detection

During the transfer phase, participants continued to benefit from the detailed feedback they received during training. As seen in Figure 4a, the detection rates of participants in the Detailed condition were higher than those of the participants in the aggregated conditions in both similar and novel scenarios, $t(158)=8.531$, $p<.001$ and $t(158)=4.231$, $p<.001$, respectively. Furthermore, there were no differences in the detection rates of participants that were trained with Detailed feedback between the Similar (mean=.80, SD=.18) and Novel (mean=.79, SD=.18) scenarios, $t(158)=.492$, p=ns. However, for participants that received aggregated feedback during training, we find higher detection rates in the Novel scenarios (mean=.62, SD=.31) compared to the Similar scenarios (mean=.49, SD=.27), $t(158)=2.829$, $p=.005$. As seen in Figure 4b, when the scenarios were similar, false alarm rates of participants in the Detailed condition were significantly lower than those of participants in the Aggregated condition, $t(158)=6.004$, $p<.001$. For the novel scenarios, however, the differences between the feedback conditions were only marginal, $t(158)=1.628$, $p=.106$. Furthermore, in the Detailed feedback condition we find an marginally significant increase in the false alarm rate when the transfer environment is novel, $t(158)=1.675$, $p=.096$. In contrast, for the Aggregated feedback condition we find a significant decrease in the false alarm rate when the transfer environment is novel, $t(158)=3.172$, $p=.002$.

5.2.2. Attributes and rule structure

Results indicated that participants in the Aggregated condition better classified the *Alert+Load* and *Alert+Operation* types in the novel environment (mean=.81, SD=.35 and mean=.50, SD=.34) compared to the similar environment (mean=.57, SD=.35 and mean=.33, SD=.34), $t(78)=3.069$, $p=.003$ and $t(78)=2.199$, $p=.031$, respectively. The hit rates of events in the form of *Alert+Load+Operation* was not influenced by the transition to novel or similar network scenario.

6. Discussion and conclusions

We investigate how feedback format during training affects the detection of novel and familiar types of attacks in cyber security scenarios. Through feedback and without an explicit definition of relevant attributes, we expected that people would learn the rules to accurately identify threats. During training, detailed feedback facilitated threat detection compared to aggregated feedback. Furthermore, detailed feedback helped identify the attributes that lead to better threat identification. Thus, detailed feedback supported the transfer of the structural knowledge, including

rules and attributes and is essential for successful training. After training with detailed feedback, participants were accurate in detecting cyber-attacks in novel and familiar scenarios compared to participants trained with aggregated feedback. The difficulty of detecting threats in the form of *Alert+Operation* in the aggregated feedback conditions replicates the findings from Ben-Asher and Gonzalez [6], and demonstrates how detailed feedback can counteract the tendency to overlook a relevant attribute in the detection process.

In contrast to a training environment that provides immediate and accurate feedback, analysts typically operate in an environment with limited and delayed feedback. Detection in a novel environment after training with aggregated feedback resembles the detection of zero-day attacks. Improvement in this type of task suggests that transfer to a novel environment triggered more careful evaluation of attributes and a more deliberate process of rule generation [18, 19]. This process depends on deep similarity, rather than on surface similarity. Developing such abilities are curial for successful detection of novel attacks.

Appropriate training may support attribute identification and rule construction by directing attention to the attributes relevant to good performance in the task. Thus, detailed feedback may help individuals acquire knowledge about the task structure. However, in addition to training with detailed and timely feedback, cyber analysts should also learn to abstract information and look more broadly at outcome feedback to improve their ability to detect novel threats.

Acknowledgements

This research was partly supported by a Multidisciplinary University Research Initiative Award (MURI; # W911NF-09-1-0525) from Army Research Office and by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] J. McHugh, Intrusion and intrusion detection. International Journal of Information Security, 1(1) (2001) 14-35.
[2] S. Jajodia, P. Liu, V. Swarup, & C. Wang, Cyber situational awareness: Issues and research. New York, NY: Springer, (2010).
[3] J. R. Goodall, W. G. Lutters, A. & Komlodi, Developing expertise for network intrusion detection. Information Technology & People, 22(2), (2009) 92-108.
[4] A. D'Amico, & K. Whitley, The real work of computer network defense analysts. In J. R. Goodall, G. Conti, & K. L. Ma (Eds.), Proceedings of the Workshop on Visualization for Computer Security. Berlin, Germany: Springer-Verlag, (2008) 19-37.
[5] C. Gonzalez, N. Ben-Asher, A. Oltramari, & C. Lebiere, Cognition and technology. In A. Kott, C. Wang, & R. Erbacher (Eds.), Cyber defense and situation awareness, (2014) 93-117.
[6] N. Ben-Asher, C. Gonzalez, Effects of cyber security knowledge on attack detection, Computers in Human Behavior (2015) 51-61.
[7] W. K. Balzer, M. E. Doherty, & R. O'Connor, Effects of cognitive feedback on performance. Psychological Bulletin, 106(3) (1989) 410-433.
[8] R. C. Haygood, & L. E. Bourne Jr, Attribute- and rule-learning aspects of conceptual behavior. Psychological Review, 72(3) (1965) 175-195.
[9] C. M. Walker & L. E. Bourne, The identification of concepts as a function of amounts of relevant and irrelevant information. The American Journal of Psychology, (1961) 410-417.
[10] J. Fodor, Concepts: A potboiler. Cognition, 50(1) (1994) 95-113.
[11] J. Feldman, Minimization of Boolean complexity in human concept learning. Nature, 407(6804) (2000) 630-633.
[12] R. L. Goldstone, D. L. Medin, & D. Gentner, Relational similarity and the nonindependence of features in similarity judgments. Cognitive Psychology, 23(2) (1991) 222-262.
[13] M. T. Chi, P. J. Feltovich, & R. Glaser, Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2) (1981) 121-152.
[14] F. Valeur, G. Vigna, C. Kruegel, & R. A. Kemmerer, Comprehensive approach to intrusion detection alert correlation. IEEE Transactions on Dependable and Secure Computing, 1(3) (2004) 146-169.
[15] S. Caltagirone, A. Pendergast, & C. Betz, The Diamond Model of Intrusion Analysis, ADA586960 (Hanover, MD: Center for Cyber Threat Intelligence and Threat Research) (2013)
[16] E. M. Hutchins, M. J. Cloppert, & R. M. Amin, Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research, 1, (2011) 80-91.
[17] K.W. Lye, & J. M. Wing, Game strategies in network security. International Journal of Information Security, 4(1-2) (2005) 71-86.
[18] V. Dutt, A. Young-Suk, N. Ben-Asher, C. & Gonzalez, Modeling the effects of base-rates on cyber threat detection performance. Proceedings of the 11th International Conference on Cognitive Modeling. Berlin, Germany (2012).
[19] T. Betsch, K. Fiedler, & J. Brinkmann, Behavioral routines in decision making: The effects of novelty in task presentation and time pressure on routine maintenance and deviation. European Journal of Social Psychology, 28(6) (1998) 861-878.